Indonesian Journal of Electrical Engineering and Computer Science
Vol. 9, No. 2, February 2018, pp. 447~459
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v9.i2.pp447-459

Educational Data Mining and Analysis of Students' Academic Performance Using WEKA

Sadiq Hussain[1], Neama Abdulaziz Dahan[2], Fadl Mutaher Ba-Alwi[3], Najoua Ribata[4]
[1]Examination Branch, Dibrugarh University, India
[2,3]Department of Computer Science, Sana'a University, Sana'a, Yemen
[4]Lirosa Laboratory, Abdelmalek Essaâdi University, Tetuan, Morocco

Article Info

Article history:

Received Sep 2, 2017
Revised Dec 25, 2017
Accepted Jan 11, 2018

Keywords:

Educational data mining
Classification algorithms
WEKA
Students' academic performance

ABSTRACT

In this competitive scenario of the educational system, the higher education institutes use data mining tools and techniques for academic improvement of the student performance and to prevent drop out. The authors collected data from three colleges of Assam, India. The data consists of socio-economic, demographic as well as academic information of three hundred students with twenty-four attributes. Four classification methods, the J48, PART, Random Forest and Bayes Network Classifiers were used. The data mining tool used was WEKA. The high influential attributes were selected using the tool. The internal assessment attribute in the continuous evaluation process makes the highest impact in the final semester results of the students in our dataset. The results showed that random forest outperforms the other classifiers based on accuracy and classifier errors. Apriori algorithm was also used to find the association rule mining among all the attributes and the best rules were also displayed.

Corresponding Author:

Sadiq Hussain,
Examination Branch, Dibrugarh University, India
Email: sadiq@dibru.ac.in

1. INTRODUCTION

Data Mining (DM) is one of the active fields in the Computer Sciences (CSs). It is a young and promising field. Due to the extensivity and the huge availability of the amounts of data and the urgent need to convert such data into useful information and knowledge, Data mining has enticed a great importance of interest in the information industry and in society as well in recent years [1]. DM focuses on the extraction of hidden knowledge from various data warehouses, data marts, and repositories. Large data becomes useless without proper utilization.

Sometimes DM can be named also Knowledge data discovery (KDD). They are similar in many things but they are really different in an essential point. DM is to find a subset Di of D that met a logical formula within the scope of Di reduced matrix. If DM cannot deduced any results from that logical formula, KDD will be found, in contrast, even if that logical formula can cover all the data as well as the possibility of the knowledge discovery. The main feature of both data mining and knowledge discovery is to derive common expressions of characteristics that are shared by all elements in a set [2]. KDD and DM have techniques that are used to extract useful information from large amount of data in the database [3, 4]. The results of applying the DM algorithms on any given or manual-generated dataset can be named the Rule Discovery [5]. There are two main types of these rules, the production rules and the association rules. According to Quinlan [6], the production rules are a common formalism for expressing knowledge in expert systems. Decision Trees rules can be also transformed into the production rules [6]. The association rules was firstly addressed to find a relationship among sales of different items from the analysis of a big data [7].There are many fields that DM has been applied in, One of them is the educational DM (EDM).

Educational data mining is an emerging field in the area of data mining. In this competitive world, the educational setting also uses data mining tools to explore and analyze student performance, predict their results to prevent drop out and focus on both good and academically poor performers, feedback for the faculties and instructors, visualization of data and to have a better assessment of learning process. The quality of education needs to be improved and educational data mining is a tool for this improvement. Modern educational institutes need data mining for their strategy and future plans. Student's performance depends on various factors like personal, social, economic and other environmental ones[8, 9]. The top-level educational institutes' authorities may utilize the outcome of the experimental results to understand the trends and behaviors in students' performance which may lead to design new pedagogical strategies [10].

There are a number of classification algorithms: Decision Tree, Neural Network, Naïve Bayes, K-Nearest neighbor, Random Forest, AdaBoost, Support Vector Machines etc. [11]. In this research, authors are going to use notably some of them for mining the academic students' performance: J48, BayesNet, PART and Random Forest classification algorithms. Apriori algorithm, as a part of the unsupervised learning and one of the most popular algorithms for association rule mining was used additionally to reveal the hidden rules from our dataset [12]. They compared each of the algorithms based on its accuracy to select the best performed algorithm for the job.

Classification is one of the predictive tasks [1] and is the most commonly used data mining technique in predicting the students' performance in educational institutes [11, 13, 14]. Several attributes were considered in our study. To find the high influence attributes, feature selection was conducted first. Feature selection removes the unnecessary attributes from the dataset to extract useful and meaningful information. It makes the mining process faster, valuable and meaningful. In the study, students' end semester percentage is selected as the dependent parameter. The percentages are categorized as 'Best', 'Very Good', 'Good', 'Pass', 'Fail'. The data mining tool used for the study was WEKA (Waikato Environment for Knowledge Analysis). WEKA is an open source tool written in Java that is widely used by the data miners [15]. WEKA implements most of the machine learning algorithms and visualizes its results as well.

The paper is organized as follows: in Section II a review of related literature is presented, Section III introduces Classifier evaluations and Error Measurement Techniques used in this research. Section IV provides Applied Data mining algorithms on the selected dataset. Section V showed experimental results, Section VI presents the Association rule mining work, and section VII concludes the work.

2. LITERATURE REVIEW

Ahmad et al [16] designed a framework to predict the academic performance of the first year bachelor students of computer science course. The dataset contained 8 years data starting from July 2006-07 to July 2013-14. The data collected contained various aspects of students' records including previous academic records, family background and demographics. Three classifiers viz. Decision Tree, Naïve Bayes and Rule-Based classifiers are applied to find the academic performance of students. The experiments showed that Rule Based classifier was the best among the other classifiers and its accuracy was found as 71.3%. The first year students' level of success was predicted by the model. Sumitha et. al. [17] developed a data model to predict student's future learning outcomes using senior students dataset. They compared the data mining classification algorithms and found that J48 algorithm was best suited for such job based on their data.

Khasanah et. al. [18] conducted a study to find that high influence attributes may be selected carefully to predict student performance. Feature selection may be used before classification for such job. The student data was from Department of Industrial Engineering Universitas Islam Indonesia. They used Bayesian Network and Decision Tree algorithms for classification and prediction of student performance. The Feature Selection methods showed that student's attendance and Grade Point Average in the first semester topped the list of features. When the accuracy rate was considered, the Bayesian Network outperformed the Decision Tree classification in their case. Ankita A Nichat et. al. [19] built classification models using decision tree and artificial neural network techniques. They used several attributes to access the strength and weakness of the students to improve the performance of the students.

Hilal Almarabeh [20] used WEKA tool to evaluate the performance of the university students. He found that the accuracy of the classifier algorithms depends upon size and nature of data. The author used Naïve Bayes, Bayesian Network, Neural Network, ID3 and J48 classification techniques. It was found that Bayesian Network outperforms the others in terms of accuracy. Amjad Abu Saa [21] worked out a qualitative model to analyze the student performance based on students' personal and social factors. The author explored theoretically various factors of the students' performance in the field of higher education.

Pedro Strecht et. al. [10] predicted students' results (pass/fail) and their grades in their work. They used classification model for the students' results and a regression model for the prediction of the grades.

They carried out the experiments using the 700 courses students' data who studied at the University of Porto. They used decision trees and SVM for classification while SVM, Random Forest, and AdaBoost.R2 were best suited for regression analysis. The classification model was able to extract useful patterns, but the models for regression were not able to beat a simple baseline. Fahim Sikder et. al. [13] used Cumulative Grade Point Average (CGPA) for prediction of students' yearly performance. The dataset used was from Bangabandhu Sheikh Mujibur Rahman Science and Technology University students' records. The authors used neural network technique for prediction and it was compared with the real CGPA of the student.

2.1 Classifier Evaluations and error measurement techniques:

The performance measures are derived from confusion matrix [22]. A confusion matrix is formed based on the four outcomes of binary classification. In binary classification, the dataset usually has two labels positive (P) and negative (N). The outcomes are true positive (TP) i.e. correct positive prediction, true negative (TN) i.e. correct negative prediction, false positive (FP) i.e. incorrect positive prediction and false negative (FN) i.e. incorrect negative prediction.

a. Sensitivity (Recall or True positive rate)
Recall is the number of correct classifications divided by the total number of positives. So,

$$R = TP/(TP + FN) = TP / P \quad (1)$$

b. Precision
Precision is the number of correct positive classifications divided by total number of positive classifications. So,

$$P = TP / (TP + FP) \quad (2)$$

c. F-score
F-score is harmonic mean of precision and recall. So,

$$F = 2PR / (P+R) \quad (3)$$

d. Accuracy [23]
Accuracy is the number of all correct classifications divided by the total numbers of cases. So,

$$Accuracy = (TP+TN) / (TP+TN+FN+FP) = (TP+TN) / (P+N) \quad (4)$$

The following section explains different error measures used for classification methods.

e. Mean Absolute Error (MAE) [24]
MAE estimates how far the predictions or forecasts differ from the actual values.

$$\mathbf{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} | \boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{x} | \quad (5)$$

where n = the number of errors, |xii – x| = the absolute errors.

f. Root Mean Square Error (RMSE) [24]
RMSE is an evaluator of the differences between the predictor values and the actual observed values.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_{obs,i} - X_{model,i})^2}{n}} \quad (6)$$

where Xobs is observed values and Xmodel is modeled values at time/place i.

g. Relative Absolute Error (RAE) [20] [15]
RAE is defined as the ratio of absolute error by the magnitude of the actual value. It is represented as below,

$$RAE = \frac{\sum_{i=1}^{n} |p_i - a_i|}{\sum_{i=n}^{n} |\bar{a} - a_i|} \quad (7)$$

where pi is the forecast value, ai is the actual value and $\bar{a}$ is the average of actual values.
h. Root Relative Squared Error (RRSE) [20] [15]
It is denoted as mean absolute error (MAE) divided by the classification model error. It can be represented as below,

$$RRSE = \sqrt{\frac{\sum_{i=1}^{n} (b_i - a_i)^2}{\sum_{i=1}^{n} (a_i - \bar{a})^2}} \quad (8)$$

Avoiding bias in the algorithms selection:

There were many studies for accessing the student academic performance and prediction of drop out of students and their job prospects [25]. The goal of such type of study was to improve the quality of education in higher educational institutes. Most of the studies consider the grade point averages (GPA) [26, 27], as their response variable and the explanatory variables are varied. In our study, we had used final semester percentage as our response variable as the grading system are not yet introduced at undergraduate level in most of the courses in Assam.

There were also various classification methods applied for student academic performance studies [16, 20]. The different studies showed that on their dataset the results found on accuracy varies. Some of the studies found that the decision trees are the best among other classification algorithms whereas some found that Bayes Network performed better than others.

The authors had applied four of the classification methods one by one until the accuracy found to be 99% in case of random forest. The first method used by the authors was Bayesian Network (BN). According to Almarabeh [20] had analyzed the performance of students' of King Saud Bin Abdulaziz University for Health Sciences. He found that BN was the best-suited classification methods. Directed acyclic graphs are used in Bayesian networks to depict the dependencies among random variables. Random variables are represented as nodes. If the nodes are connected by an arc, then these variables are dependent on each other. BN has been used for performing bi-directional inference since 1980. It is also used for reasoning under uncertainty.

The authors then tried the rule-based classification techniques available as PART in WEKA. Ahmad et al [16] also used this technique for classification and found that it was the best technique for student academic performance assessment among Naïve Bayes, decision trees, and rule-based classifiers. PART is rules-based classifier which combines separate and conquer method with divide and conquer strategy. This classification method builds a partial tree with the available set of records. It then creates a rule from the tree. After discarding the decision tree and deleting records covered by the rule, it again builds the partial decision tree in an iterative manner.

The authors then used the decision tree classification method. Patil et al [28] established that decision tree algorithm performs better than Naïve Bayes methods. The advantage of using decision tree classifier is that the tree can be visualized, understood and interpreted easily by the users [29]. The tree performs well in case of both numerical and categorical variables. The decision tree has a tree-like structure start with root node and ends with leaf attributes. So, it is one of the powerful as well as popular classifiers. WEKA implements C4.5 decision tree using J48 classification method.

The authors used random forest classifier as their next attempt. Random forests (RF) [11] reduce overfitting, bias, and variance. So, RF is more accurate and robust. RF works on bagging algorithm. RF replaces data to construct the tree and the partition is not done on the same important variable as the explanatory variables are bootstrapped. RF creates lots of individual decision trees from the training set. It is good at predicting the target values.

4. APPLYING DATA MINING ALGORITHMS TO THE SELECTED DATASET

The dataset contained 300 instances with 24 attributes. The proposed framework is shown in Figure 1 below.

4.1. Data Preprocessing phase

The data for this research was collected from three different colleges, those are Duliajan College, Doomdooma College and Digboi College of Assam, India. Initially, data of twenty-four attributes were collected. As the attribute name of the student does not carry any significance, we removed it from the list of the attributes. The attribute "marks in practical paper" was also removed at the pre-processing phase, because of the interesting number of the missing values. Finally, twenty-two attributes were selected after data cleaning. Table-1 shows the selected attributes with their possible values.

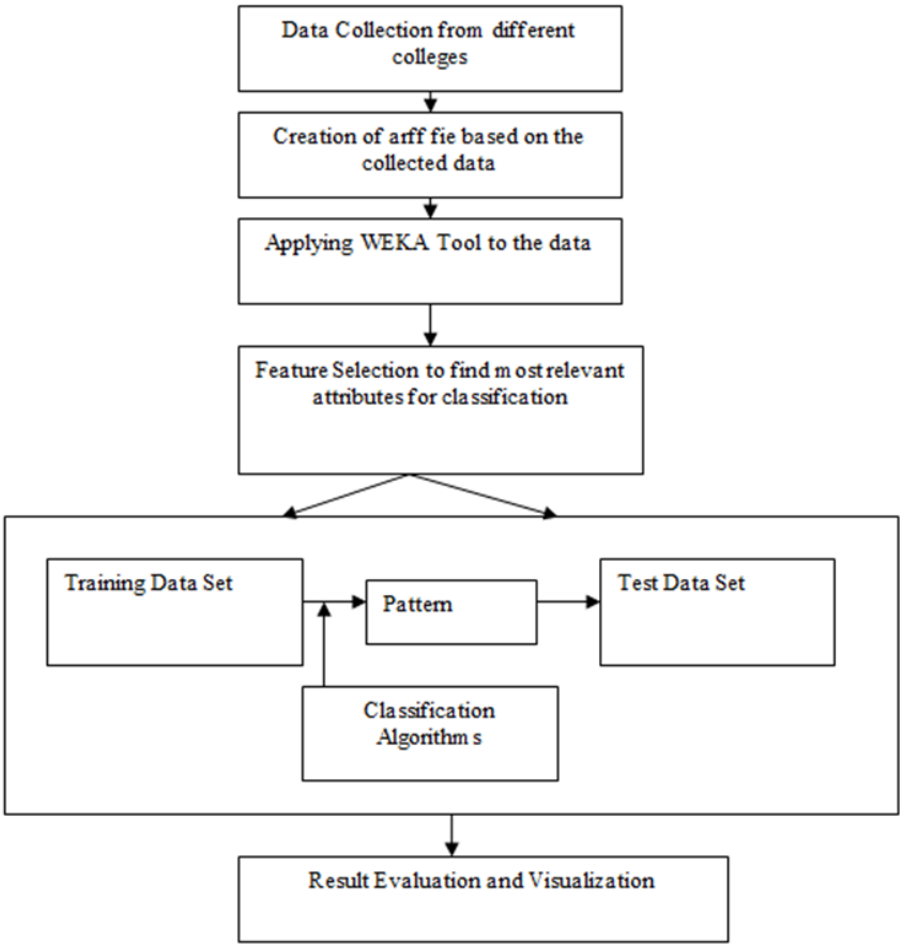

Figure 1: Framework for Students' Academic Performance Classification

Table 1: Dataset Description

Attribute	Description	Values
GE	Gender	(Male, Female)
CST	Caste	(General,SC,ST,OBC,MOBC)
TNP	Class X Percentage	(Best, Very Good, Good, Pass, Fail) If percentage >=80 then Best If percentage >= 60 but less than 80 then Very Good If percentage >= 45 but less than 60 then Good If Percentage >= 30 but less than 45 then Pass If Percentage < 30 then Fail
TWP	Class XII Percentage	(Best, Very Good, Good, Pass, Fail) Same as TNP
IAP	Internal Assessment Percentage	(Best, Very Good, Good, Pass, Fail) Same as TNP
ESP	End Semester Percentage	(Best, Very Good, Good, Pass, Fail) Same as TNP
ARR	Whether the student has back or arrear papers	(Yes, No)
MS	Marital Status	(Married, Unmarried)
LS	Lived in Town or Village	(Town, Village)
AS	Admission Category	(Free, Paid)
FMI	Family Monthly Income (in INR)	(Very High, High, Above Medium, Medium, Low) If FMI >= 30000 then Very High If FMI >= 20000 but less than 30000 then High If FMI >= 10000 but less than 20000 then Above Medium If FMI >= 5000 but less than 10000 then Medium If FMI is less than 5000 then Low The figures are expressed in INR.
FS	Family Size	(Large, Average, Small) If FS > 12 then Large If FS >= 6 but less than 12 then Average If FS < 6 then Small
FQ	Father Qualification	(IL, UM, 10, 12 , Degree, PG) IL= Illiterate UM= Under Class X
MQ	Mother Qualification	(IL, UM, 10, 12 , Degree, PG) IL= Illiterate UM= Under Class X
FO	Father Occupation	(Service, Business, Retired, Farmer, Others)
MO	Mother Occupation	(Service, Business, Retired, Farmer, Others)
NF	Number of Friends	(Large, Average, Small) Same as Family Size
SH	Study Hours	(Good, Average, Poor) >= 6 hours Good >= 4 hours Average < 2 hours Poor
SS	Student School attended at Class X level	(Govt., Private)
ME	Medium	(Eng,Asm,Hin,Ben)
TT	Home to College Travel Time	(Large, Average, Small) >= 2 hours Large >=1 hours Average < 1 hour Small
ATD	Class Attendance Percentage	(Good, Average, Poor) If percentage >= 80 then Good If percentage >= 60 but less than 80 then Average If Percentage < 60 then poor

Descriptions of some of the attributes of the dataset

CST: It is caste of the student. The possible values of this attribute are 'G' (General category or unreserved category), 'SC' (Schedule Caste category), 'ST' (Schedule Tribe Category), 'OBC' (Other Backward Classes), 'MOBC' (Minorities and other backward classes) students. These categories are based on the Indian Constitution.

TNP: It is the percentage attained by the student in Class X. The examination is called HSLC Examination in Assam, India. The authors had categorized the results as Best, Very Good, Good, Pass, Fail. The 'Best' is called when the student secured more than or equal to 80% (it is termed as Star percentage), 'Very Good' is labeled as when the student secures more than or equal to 60% but less than 80% (more than or equal to 60% is always termed as First Division or Class in most of the examinations), 'Good' is termed as when the student secures more than or equal to 45% but less than 60% (in most of the Universities in Assam it is called as Second Division or class), 'Pass' is called when the student got less than or equal to 30% but less than 45%. It is termed as 'Fail' when the student secured less than 30%. The same is true for TWP (Class XI percentage secured by the student), IAP (Internal Assessment percentage secured by the student at Degree level (10+2+3)) and ESP (End Semester Examination percentage secured by the student at Degree level). ESP is the response variable.

IAP (Internal Assessment percentage secured by the student at Degree level (10+2+3)): Internal Assessment is part of continuous evaluation. It comprises of sessional examinations, surprise tests, assignments, field work, quizzes etc. It is categorized as the same way as TNP,TWP and ESP.
ARR: It is categorized as 'Yes' or 'No'. This attribute collected the data based on the fact that whether the student had any failed paper in any of the previous semesters.
ME: It is categorized as 'Eng' (English), 'Asm' (Assamese), 'Hin' (Hindi) and Ben ('Bengali'). Assamese, Hindi and Bengali are the modern Indian languages. It is the language or medium of instructions for the students in which languages they were being taught or appeared in an examination.
FQ: The possible values of this attribute are 'Il' (illiterate), 'Um' (Under class X level), '10' (Passed Class X Examination), '12' (Passed Class XII Examination), 'Degree' (Passed Bachelor of Arts or Science or Commerce Examination), 'PG' (passed Masters of Arts or Science or Commerce Examination). It is the educational qualification of father of student. MQ stands for mother qualification. The possible values of this attribute are same as father qualification.

4.2 Feature Selection

Using Weka, the feature selection discovers the most influential attributes using correlation-based attribute evaluation, gain-ratio attribute evaluation, information-gain attribute evaluation, relief attribute evaluation, symmetrical uncertainty attribute evaluation. Correlation-based attribute evaluation is a greedy search method while others are rank search methods [18].

Using these feature selection methods, total eleven attributes were found to be highly influential. The selected attributes are shown as bold in Table 2. They were used for classification and other attributes were removed. The end semester percentage (esp) is the response variable. Figure 2 shows the data in the arff format.

Table 2: Attribute Selection using feature selection methods

Feature Selection Method	High Influence Attributes
Correlation-based Attribute Evaluation	arr, iap,tnp,as,twp,sh,me,fs,nf, atd,fo,fmi,fq,tt,ss
Gain-Ratio Attribute Evaluation	iap,ms,arr,tnp,twp,as,me,sh,atd,fmi,fq,nf,fo,mq,fs
Information-Gain Attribute Evaluation	iap,tnp,twp,arr,fmi,as,fq,me,atd,sh,fo,mq,nf,cst,tt
Relief Attribute Evaluation	iap,tnp,arr,tnp,nf,as,atd,me,fo,sh,fmi,fs,ls,ge,tt
Symmetrical Uncertainty Attribute	iap,tnp,twp,arr,as,me,fmi,atd,sh,fq,fo,mq,nf,fs,tt

```
@RELATION Sapfile1

@ATTRIBUTE ge  {M,F}
@ATTRIBUTE cst {G,ST,SC,OBC,MOBC}
@ATTRIBUTE tnp {Best,Vg,Good,Pass,Fail}
@ATTRIBUTE twp {Best,Vg,Good,Pass,Fail}
@ATTRIBUTE iap {Best,Vg,Good,Pass,Fail}
@ATTRIBUTE esp {Best,Vg,Good,Pass,Fail}
@ATTRIBUTE arr {Y,N}
@ATTRIBUTE ms  {Married,Unmarried}
@ATTRIBUTE ls  {T,V}
@ATTRIBUTE as  {Free,Paid}
@ATTRIBUTE fmi {Vh,High,Am,Medium,Low}
@ATTRIBUTE fs  {Large,Average,Small}
@ATTRIBUTE fq  {Il,Um,10,12,Degree,Pg}
@ATTRIBUTE mq  {Il,Um,10,12,Degree,Pg}
@ATTRIBUTE fo  {Service,Business,Retired,Farmer,Others}
@ATTRIBUTE mo  {Service,Business,Retired,Housewife,Others}
@ATTRIBUTE nf  {Large,Average,Small}
@ATTRIBUTE sh  {Good,Average,Poor}
@ATTRIBUTE ss  {Govt,Private}
@ATTRIBUTE me  {Eng,Asm,Hin,Ben}
@ATTRIBUTE tt  {Large,Average,Small}
@ATTRIBUTE atd {Good,Average,Poor}

@DATA
F,G,Good,Good,Vg,Good,Y,Unmarried,V,Paid,Medium,Average,Um,10,Farmer,Housewife,Large,Poor,Govt,.
```

Figure 2: Data File in arff format

4.3 Specifying the selected algorithms

After feature selection, the classification algorithms were applied. There are various classification methods: Decision Tree, Neural Network, Naïve Bayes, K-Nearest neighbor, Random Forest, AdaBoost, Support Vector Machines etc. [13]. The authors used specific algorithms, for mining the academic performance of the students, those are found in the WEKA program: J48, PART, BayesNet and Random Forest classification algorithms. According to the WEKA algorithms specification [30]: J48 is an algorithm that generates a pruned or unpruned C4.5 decision tree. PART is an algorithm that uses divide-and-conquer mechanism to build a partial C4.5 decision tree in each iteration, i.e. it generates a PART decision list, and makes the best leaf into a rule. BayesNet produces random instances based on a Bayes network that uses various search algorithms and quality measures. It also offers data structures (network structure, conditional probability distributions, etc.) and facilities public to Bayes Network learning algorithms. Random Forest is a group of unpruned classification or regression trees that are created using bootstrap examples of the training data and random feature selection in tree induction that is finally constructing a forest of random trees [30, 31]. Then the authors compared each of the algorithms based on its accuracy to select the best-performed algorithm for the job.

5. EXPERIMENTS AND RESULTS

5.1 Classification Results:

The stage is set for the experiments. WEKA has various classification algorithms. The authors had used J48, BayesNet, PART and Random Forest classification methods available in WEKA. These methods are supervised learning algorithms which use the training data to test the correctness of testing data [20]. Figure 4 shows the comparison between these four classifiers.

J48 Classifier: This classifier is used for generating decision tree based on C4.5 algorithm. Ross Quinlan developed this algorithm [20]. Its performance is shown in figure 6.

BayesNet Classifier: This classifier delivers higher accuracy on large database. It also makes the computational time less than better speed. Bayesian Network uses conditional dependencies using direct graph [20].

Random Forest Classifier: This classifier used bootstrap sampling method on the training dataset to construct many unpruned classification trees. In the testing phase, the mean of all unpruned classification trees for a randomly selected feature provides the final predicted output [32]. Its performance is shown in figure 7 and 8.

PART Classifier: This rue learning classifier combines the divide-and-conquer strategy with separate-and conquers strategy. It builds a partial decision tree on the current set of instances and creates a rule from the decision tree [33].

There are 300 student records from three different colleges with 12 selected attributes. Table 3 shows the performance of the 4 classification methods based on their accuracy.

Table 3: Comparison of different classifiers based on accuracy.

Classifiers	Accuracy	Correctly Classified Instances	Incorrectly Classified Instances
Random Forest	99%	297	3
PART	74.33%	223	77
J48	73%	219	81
BayesNet	65.33%	196	104

Based on the accuracy of the four classifiers, the Random Forest has more correctly classified instances than other classification methods. Its accuracy percentage is 99%. Figure 4 and 5 shows that the Random Forest Classifier has the minimum errors in terms of Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE) when compared with other methods.

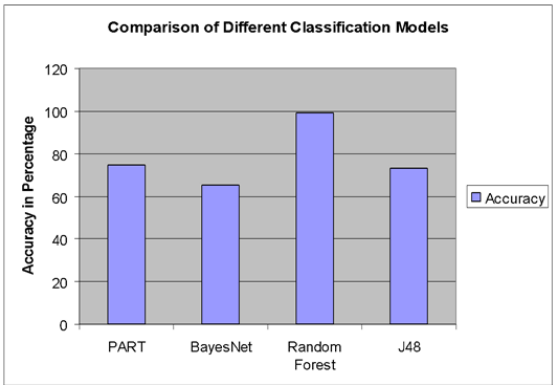

Figure 3. Comparison of Classifiers

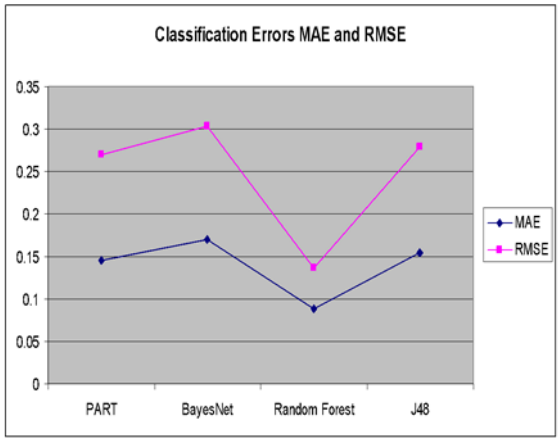

Figure 4: MAE and RMSE Metrics

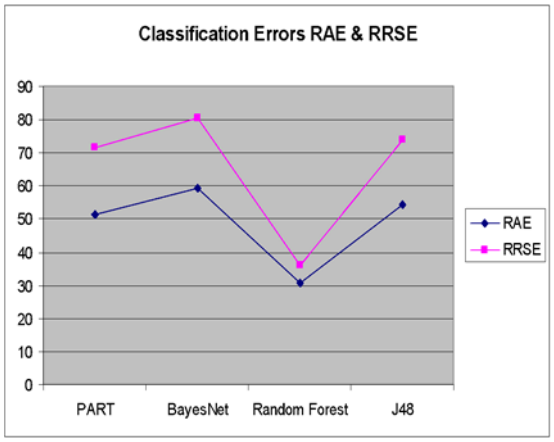

Figure 5: RAE and RRSE Metrics

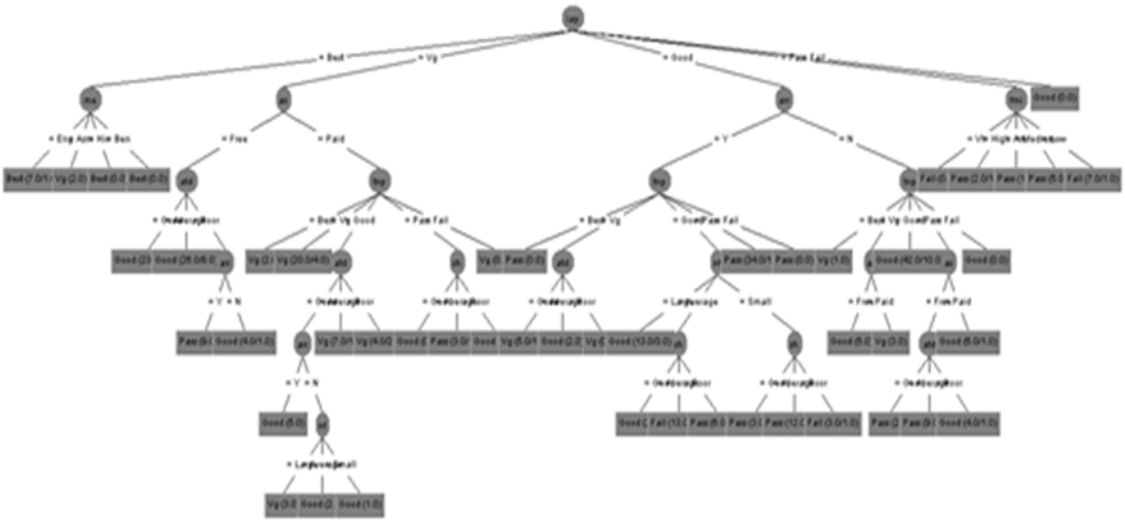

Figure 6. J48 Tree Visualization

The Kappa statistic value is 0.9859 which shows that the model is statistically significant.
The significance is rather high according to this value. So, this model may be used for the prediction of final semester percentage of the student.

The authors had also compared the random forest classifier with feature selection and without feature selection. The random forest classifier with feature selection outperforms the other. Table 4 shows the comparison.

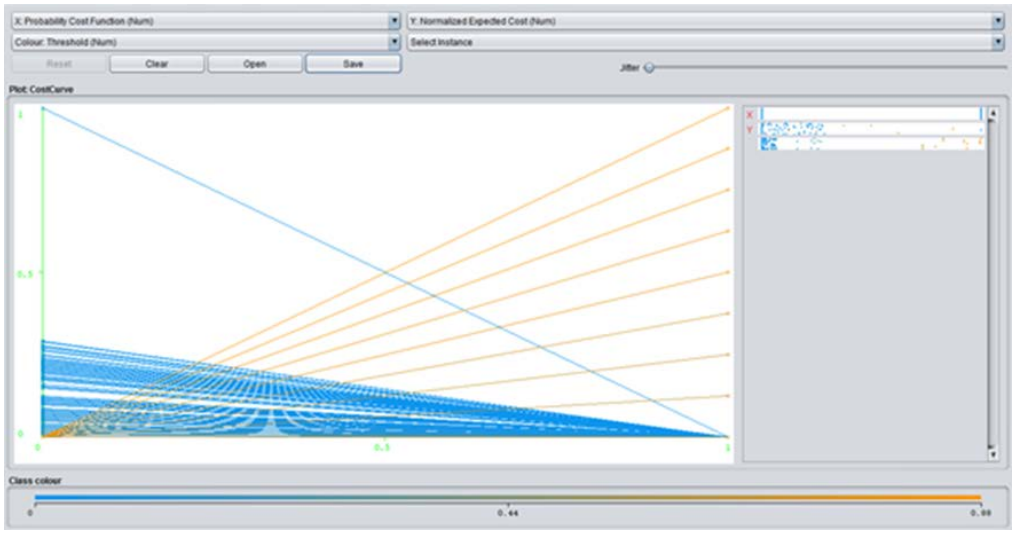

Figure 7: Random Forest Visualization of Cost Curve of 'Best' Class of 'end semester percentage' attribute

Table 4: Comparison of Random Forest Classifier with and without selected attributes

Classifiers	Accuracy	Correctly Classified Instances	Incorrectly Classified Instances
Random Forest With 12 selected attributes	99%	297	3
Random Forest With all the attributes	84.33%	233	67

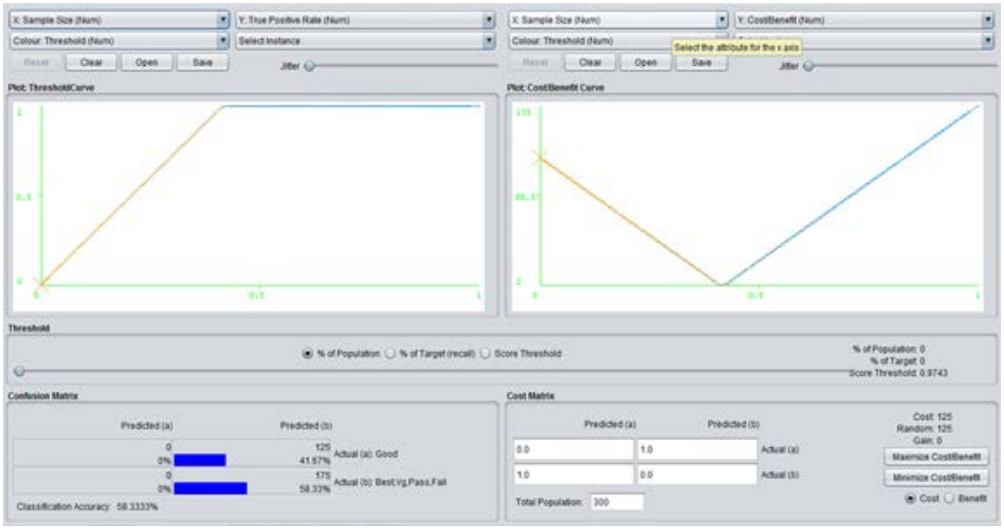

Figure 8. Random Forest Visualization of Cost/Benefit Analysis for 'Good' Class of 'end semester percentage' attribute

5.1 Association Rule results

Association rules are used for analyzing the data to uncover the frequent if/then patterns. The most important relationships are identified by using support and confidence criteria. The association rule comprises of two parts. They are antecedent (if part) and a consequent (then part). The Apriori algorithm is most frequently used algorithm to find the correlation based data mining works [12]. Using WEKA, we had applied the Apriori algorithm on our datasets. The Minimum support was 0.6 (180 instances), minimum metric (confidence) was 0.9 and number of cycles performed were 8. We had found the best rules as shown below:

1. ls=V 240 ==> ms=Unmarried 240 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.8)
2. ls=V mo=Housewife 213 ==> ms=Unmarried 213 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.71)
3. fs=Small 202 ==> ms=Unmarried 202 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.67)
4. as=Free 191 ==> ms=Unmarried 191 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.64)
5. fs=Small mo=Housewife 182 ==> ms=Unmarried 182 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.61)
6. ls=V ss=Govt 181 ==> ms=Unmarried 181 <conf:(1)> lift:(1) lev:(0) [0] conv:(0.6)
7. mo=Housewife 269 ==> ms=Unmarried 268 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.45)
8. ss=Govt 221 ==> ms=Unmarried 220 <conf:(1)> lift:(1) lev:(-0) [0] conv:(0.37)
9. mo=Housewife ss=Govt 200 ==> ms=Unmarried 199 <conf:(0.99)> lift:(1) lev:(-0) [0] conv:(0.33)
10. me=Asm 193 ==> ms=Unmarried 192 <conf:(0.99)> lift:(1) lev:(-0) [0] conv:(0.32)

The experiment was again performed with the selected attributes. This time the Minimum support was 0.1 (30 instances), minimum metric (confidence) was 0.9 and number of cycles performed were 18. The authors had found the best rules as shown below:

1. esp=Fail 32 ==> arr=Y 32 <conf:(1)> lift:(1.97) lev:(0.05) [21] conv:(15.79)
2. fmi=Low fo=Farmer me=Asm 32 ==> as=Free 32 <conf:(1)> lift:(1.57) lev:(0.04) [15] conv:(11.63)
3. arr=Y fo=Farmer nf=Average 31 ==> as=Free 31 <conf:(1)> lift:(1.57) lev:(0.04) [15] conv:(11.26)
4. twp=Good iap=Good arr=Y fo=Farmer 32 ==> me=Asm 31 <conf:(0.97)> lift:(1.51) lev:(0.03) [33] conv:(5.71)
5. tnp=Good fmi=Low me=Asm 31 ==> as=Free 30 <conf:(0.97)> lift:(1.52) lev:(0.03) [33] conv:(5.63)
6. arr=Y nf=Average me=Asm 50 ==> as=Free 48 <conf:(0.96)> lift:(1.51) lev:(0.05) [13] conv:(6.06)
7. twp=Good iap=Good as=Free fo=Farmer 44 ==> me=Asm 42 <conf:(0.95)> lift:(1.48) lev:(0.05) [14] conv:(5.23)
8. fmi=Low fo=Farmer 43 ==> as=Free 41 <conf:(0.95)> lift:(1.5) lev:(0.05) [14] conv:(5.21)
9. iap=Good arr=Y fo=Farmer 43 ==> as=Free 41 <conf:(0.95)> lift:(1.5) lev:(0.05) [14] conv:(5.21)
10. iap=Good arr=Y fo=Farmer me=Asm 40 ==> as=Free 38 <conf:(0.95)> lift:(1.49) lev:(0.04) [18] conv:(4.84)

6. CONCLUSION AND FUTURE WORK

The students' academic performance was evaluated based on academic and personal data collected from 3 different colleges from Assam, India. The total number of records were 300 with 24 attributes. Two attributes were dropped in the phase of data cleaning. Using feature selection, 12 highly influential attributes were selected. After that four different classification algorithms were used. They were J48, PART, BayesNet and Random Forest. The data mining tool used in the experiment was WEKA 3.8. Based on the accuracy and the classification errors one may conclude that the Random Forest Classification method was the most suited algorithm for the dataset. The Apriori algorithm was applied to the dataset using WEKA to find some of the best rules. The data may be extended to collect some of the extra-curricular aspects and technical skills of the students and mined with different classification algorithms to predict the student performance as future work. The authors also interested in working in future on data of students assessments for each course trying to know what kind of student succeed on what kind of courses. It may define what kinds of courses are adapted for every cluster of students model who shares the same characteristics. It may also provide various multidimensional summary reports and redefine pedagogical learning paths.

ACKNOWLEDGEMENTS

The authors acknowledge the Principals of Digboi, Doomdooma and Duliajan Colleges for collecting the student data and to analyze the data to get desired results. The authors would also like to

acknowledge Prof. Alak Kr. Buragohain, Vice-Chancellor of Dibrugarh University and Prof. G.C. Hazarika, Department of Mathematics, Dibrugarh University for their inspiring words and guidance.

REFERENCES

1. Han, J., J. Pei, and M. Kamber, *Data mining: concepts and techniques*. 2 ed. 2011: Elsevier.
2. Ohsuga, S., *Difference Between Data Mining And Knowledge Discovery --A View To Discovery From Knowledge-Processing*, in *Granular Computing, 2005 IEEE International Conference on.* 2005, IEEE. p. 6.
3. Ba-Alwi, F.M. and H.M. Hintaya, *Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach*. International Journal of Scientific & Engineering Research, 2013. **4**(8): p. 6.
4. Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, *From Data Mining to Knowledge Discovery in Databases.* American Association for Artificial Intelligence Magazine, 1996. **17**: p. 18.
5. Bhatnagar, V., A.S. Al-Hegami, and N. Kumar, *A Hybrid Approach for Quantification of Novelty in Rule Discovery.* International Journal of Computer, Electrical, Automation, Control and Information Engineering, 2007. **1**(4): p. 4.
6. Quinlan, J.R., *Generating Production Rules From Decision Trees.* ijcai, 1987. **87**: p. 4.
7. Ba-Alwi, F.M., *Discovery of novel association rules based on genetic algorithms.* British Journal of Mathematics & Computer Science, 2014. **4**(23): p. 17.
8. Hijazi, S.T. and S.M.M.R. Naqvi, *Factors Affecting Students' Performance, A Case Of Private Colleges.* Bangladesh e-Journal of Sociology, 2006. **3**(1): p. 10.
9. Bhardwaj, B.K. and S. Pal, *Data Mining: A prediction for performance improvement using classification.* International Journal of Computer Science and Information Security, 2011. **9**(4): p. 5.
10. Strecht, P., et al., *A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance.* Proceedings of the 8th International Conference on Educational Data Mining, 2015: p. 3.
11. Dekker, G.W., M. Pechenizkiy, and J.M. Vleeshouwers, *Predicting students drop out: A case study.* EDM '09-Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, 2009. **2**: p. 10.
12. Shrivastava, A.K. and R.N. Panda, *Implementation of Apriori Algorithm using WEKA.* KIET International Journal of Intelligent Computing and Informatics, 2014. **1**(1): p. 4.
13. Sikder, M.F., M.J. Uddin, and S. Halder, *Predicting Students Yearly Performance using Neural Network: A Case Study of BSMRSTU.* 5th International Conference on Informatics, Electronics and Vision (ICIEV), 2016. **5**: p. 6.
14. Millán, E., T. Loboda, and J.L. Pérez-de-la-Cruz, *Bayesian networks for student model engineerin.* Computers and Education. Elsevier Ltd, 2010. **55**(4): p. 20.
15. Kabakchieva, D., *Predicting Student Performance by Using Data Mining Methods for Classification.* Cybernatics and Information Technologies, 2013. **13**(1): p. 12.
16. Ahmad, F., N.H. Ismail, and A. Abdulaziz, *The Prediction of Students' Academic Performance Using Classification Data Mining Techniques.* Applied Mathematical Sciences, 2015. **9**(129): p. 12.
17. Sumitha, R. and E.S. Vinothkumar, *Prediction of Students Outcome Using Data Mining Techniques.* International Journal of Scientific Engineering and Applied Science (IJSEAS), 2016. **2**(6): p. 8.
18. Khasanah, A.U. and Harwati, *A Comparative Study to Predict Student's Performance Using Educational Data Mining Techniques.* IOP Conf. Series: Materials Science and Engineering, 2017. **215**(012036): p. 7.
19. Nichat, A.A. and D.A.B. Raut, *Analysis of Student Performance Using Data Mining Technique.* International Journal of Innovative Research in Computer and Communication Engineering, 2017. **2007**(An ISO 3297): p. 5.
20. Almarabeh, H., *Analysis of Students' Performance by Using Different Data Mining Classifiers.* I.J. Modern Education and Computer Science, 2017. **9**(8): p. 9-15.
21. Saa, A.A., *Educational Data Mining & Students' Performance Prediction.* (IJACSA) International Journal of Advanced Computer Science and Applications, 2016. **7**(5): p. 9.
22. Willmott, C.J. and K. Matsuura, *Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance.* Climate research, 2005. **30**(1): p. 4.
23. J.Hyndman, R. and A. B.Koehler, *Another look at measures of forecast accuracy.* International Journal of Forecasting, 2006. **22**(4): p. 10.
24. Jr, R.G.P., O. Thontteh, and H. Chen, *Components of information for multiple resolution comparison between maps that share a real variable.* Pontius, Robert Gilmore, Olufunmilayo Thontteh, and Hao Chen. "Components of information for multiple resolution comparison between maps that share a real variable." Environmental and Ecological Statistics, 2008. **15**(2): p. 32.
25. Roth, P.L., et al., *(1996). Meta-analyzing the relationship between grades and job performance.* Journal of Applied Psychology, 1996. **81**: p. 8.
26. Kuncel, N.R., S.A. Hezlett, and D.S. Ones, *Academic performance, career potential, creativity, and job performance: Can one construct predict them all?* . Journal of Personality and Social Psychology, 2004. **86**(1): p. 13.
27. Kuncel, N.R., et al., *A meta-analysis of the Pharmacy College Admission Test (PCAT) and grade predictors of pharmacy student success.* American Journal of Pharmaceutical Education, 2005. **69**(3): p. 8.
28. Patil, T. and S.S. Sherekar, *Performance Analysis of Naïve Bayes and J4.8 Classification Algorithm for data classification.* International Journal of Computer Science and Applications, 2013. **6**(2): p. 5.
29. Anuradha, C. and T. Velmurugan, *A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance.* Indian Journal of Science and Technology, 2015. **8**(15): p. 12.
30. Waikato, T.U.o. *WEKA documents*.

31. Svetnik, V., et al., *Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling.* Journal of chemical information and computer sciences, 2003. **43**(6): p. 12.
32. Agrawal, S., S.K. Vishwakarma, and A.K. Sharma, *Using Data Mining Classifier for Predicting Student's Performance in UG level.* International Journal of Computer Applications, 2017. **172**(8): p. 6.
33. Tan, P.-N., M. Steinbach, and V. Kumar, *CHAPER 5: Rule-based Classifiers*, in *Introduction to Data Mining*, A. Nordman, Editor. 2005, PEARSON EDUCATION: US.